PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Iwona Kasprzyk
Akademia Ekonomiczna w Katowicach

MODELE KLAS UKRYTYCH DLA TABLIC KONTYNGENCJI Z BRAKUJĄCYMI DANYMI

1. Wstęp

Brakujące dane są powszechnym problemem w wielu typach analiz danych. Występowanie brakujących danych może mieć różne przyczyny. Często respondenci pomijają celowo niektóre pytania w ankiecie, uważając je za zbyt osobiste (np. dochody badanego). Innym razem zdarza się, że respondent nie ma wystarczającej wiedzy, aby odpowiedzieć na wszystkie pytania zawarte w ankiecie.

Metody analizy danych niepełnych można podzielić ogólnie na dwie grupy: te, które uwzględniają mechanizm powstawania brakujących danych, oraz te, które tego mechanizmu nie uwzględniają. W tym artykule zostaną pokazane trzy podstawowe mechanizmy powstawania brakujących danych: MCAR, MAR oraz NMAR.

Poniżej zostaną przedstawione wyniki badania ankietowego preferencji użytkowników telefonów komórkowych. Potrzebne obliczenia wykonano za pomocą programu LEM.

2. Modele klas ukrytych dla danych pełnych

Modele klas ukrytych pozwalają na analizę danych mierzonych na najsłabszych skalach pomiaru. Punktem wyjścia do analizy klas ukrytych jest tablica kontyngencji. W tabeli 1 przedstawiono przykład wielowymiarowej tablicy kontyngencji dla czterech zmiennych obserwowalnych A, B,C oraz D.

Do modelowania klas ukrytych używamy dwu podejść: probabilistycznego oraz log-liniowego. Pierwsze z nich zostało zaproponowane przez Lazarsfelda w latach

pięćdziesiątych, a następnie rozwijane przez Goodmana [1974]. Według tego ujęcia, model z jedną zmienną ukrytą (X) oraz czterema zmiennymi obserwowalnymi (A,B,C,D) dla prezentowanego przykładu można zapisać za pomocą równania:

$$\pi_{abcd}^{ABCD} = \sum_{x=1}^{X} \pi_{abcdx}^{ABCDX} , \tag{1}$$

gdzie: π_{abcdx}^{ABCDX} – prawdopodobieństwo warunkowe tego, że kategorie a, b, c oraz d zmiennych A, B, C, D znajdą się w klasie ukrytej x.

Tabela 1. Czterowymiarowa tablica kontyngencji dla zmiennych A, B, C, D

			D1	D2
A1	B1	C1		
		C2		
	B2	C1		
		C2		
	B3	C1		
		C2		
A2	B1	C1		
		C2		
	B2	C1		
		C2		
	B3	C1		
		C2		
A3	B1	C1		
		C2		
	B2	C1		
		C2		
	B3	C1		
		C2		

Źródło: opracowanie własne.

Wykorzystanie wzoru (1) wymaga spełnienia założenia o lokalnej niezależności:

$$\pi_{abcdx}^{ABCDX} = \pi_{x}^{X} \pi_{ax}^{A\backslash X} \pi_{bx}^{B\backslash X} \pi_{cx}^{C\backslash X} , \tag{2}$$

gdzie: π_{x}^{X} – prawdopodobieństwo tego, że dla danej obserwacji zmienna ukryta X przybiera wartość x,

$\pi_{ax}^{A\backslash X}$ – prawdopodobieństwo warunkowe tego, że kategoria a zmiennej A znajdzie się w klasie ukrytej x.

Prawdopodobieństwa po prawej stronie równania (2) wymagają spełnienia założenia:

$$\sum_{x=1}^{X} \pi_{x}^{X} = \sum_{a=1}^{A} \pi_{ax}^{A\backslash X} = \sum_{b=1}^{B} \pi_{bx}^{B\backslash X} = \sum_{c=1}^{C} \pi_{cx}^{C\backslash X} = 1 . \tag{3}$$

Podejście log-liniowe do modelowania klas ukrytych zostało zaproponowane przez Habermana [1974]. Do modelu log-liniowego została wprowadzona zmienna

ukryta X. Taki model zawiera główne efekty zmiennych obserwowalnych, zmiennej ukrytej X oraz efekty interakcji pomiędzy zmienną ukrytą X a zmiennymi obserwowalnymi A, B, C oraz D:

$$\log(F_{xabcd}) = u + u_x^X + u_a^A + u_b^B + u_c^C + u_d^D + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} + u_{xd}^{XD}, \tag{4}$$

gdzie: $F_{xabcd} = n\pi_{abcdx}^{ABCDX}$,

u – średnia logarytmów częstości oczekiwanych,

u_x^X – główny efekt zmiennej X ze względu na klasę x,

u_a^A – główny efekt zmiennej A,

u_{xa}^{XA} – efekt interakcji kategorii a zmiennej A oraz klasy x zmiennej ukrytej X.

W analogiczny sposób definiowane są pozostałe efekty główne oraz efekty interakcji.

Parametry modelu (4), tj. $\{u_x^X\},\{u_a^A\},\{u_b^B\},\{u_c^C\},\{u_d^D\},\{u_{xa}^{XA}\},\{u_{xb}^{XB}\},\{u_{xc}^{XC}\},\{u_{xd}^{XD}\}$ spełniają następujące warunki:

$$\begin{aligned}&\sum_{x=1}^{X} u_t^X = \sum_{a=1}^{A} u_a^A = \sum_{b=1}^{B} u_b^B = \sum_{c=1}^{C} u_c^C = \sum_{d=1}^{D} u_d^D = \sum_{a=1}^{A} u_{xa}^{XA} = \sum_{x=1}^{X} u_{xa}^{XA} = \\ &= \sum_{b=1}^{B} u_{xb}^{XB} = \sum_{x=1}^{X} u_{xb}^{XB} = \sum_{c=1}^{C} u_{xc}^{XC} = \sum_{x=1}^{X} u_{xc}^{XC} = \sum_{d=1}^{D} u_{xd}^{XD} = \sum_{x=1}^{X} u_{xd}^{XD} = 0.\end{aligned} \tag{5}$$

W swej pracy [1979] Haberman wykazał, że istnieje związek pomiędzy parametrami dwóch różnych od siebie podejść do modelowania klas ukrytych. Na przykład prawdopodobieństwo warunkowe tego, że kategoria „a" zmiennej A znajdzie się w klasie ukrytej „x" można zapisać:

$$\pi_{ax}^{A\backslash X} = \frac{\exp(u_a^A + u_{xa}^{XA})}{\sum_{i=1}^{I} \exp(u_{ia}^A + u_{xa}^{XA})}. \tag{6}$$

Za pomocą równania (6) można powrócić do podejścia probabilistycznego zaproponowanego przez Lazarsfelda [1968], które to przedstawia model w sposób łatwiejszy do interpretacji.

Podstawowym algorytmem stosowanym w modelowaniu klas ukrytych jest algorytm EM. Jest to główny algorytm, który został zaimplementowany w programie LEM. Składa się on z dwóch kroków: w pierwszym kroku „E" (ang. *expectation*) wyznacza się częstości oczekiwane:

$$\hat{n}_{xabcd} = n_{abcd}\hat{\pi}_{x/abcd}. \tag{7}$$

W następnym kroku „M" (ang. *maximization*) funkcja największej wiarygodności jest maksymalizowana w celu uzyskania ostatecznych wartości estymowanych parametrów:

$$\log(L_{(\pi)}) = \sum_{xabcd} \hat{n}_{xabcd} \log(\hat{\pi}_{xabcd}). \tag{8}$$

3. Metody analizy danych jakościowych ze zmiennymi z brakującymi danymi

Do estymacji parametrów modelu log-liniowego dla niekompletnych danych można zastosować dwie metody. Pierwsza z nich, zaproponowana przez Fuchsa [1982], nie pozwalała na analizę modeli log-liniowych ze zmiennymi ukrytymi. Hagenaars [1990] zmodyfikował podejście Fuchsa, wprowadzając do modelu log-liniowego zmienne ukryte.

Przypuśćmy, że mamy tablicę kontyngencji z czterema zmiennymi *A*, *B*, *C* oraz *D*. Do programu LEM wymagane jest wprowadzenie podgrup. Niech podgrupa *ABCD* oznacza, że wszystkie zmienne są obserwowalne, czyli dla żadnej z tych zmiennych nie występują brakujące dane. Podgrupa *ABC* będzie oznaczać, że zmienna *D* jest zmienną, w której występują brakujące dane. W przypadku podgrupy *ABD*, zmienne *A*, *B* oraz *D* są obserwowalne. Dla podgrupy *AB* zmienne *C* oraz *D* są zmiennymi z brakującymi danymi. Uwzględniając te informacje, krok *E* w algorytmie *EM* można zapisać:

$$\hat{n}_{xabcd} = n_{abcd}\hat{\pi}_{x/abcd} + n_{abc}\hat{\pi}_{xd/abc} + n_{abd}\hat{\pi}_{xc/abd} + n_{ab}\hat{\pi}_{xcd/ab}. \tag{9}$$

Druga metoda wprowadzona przez Faya [1986] pozwala na estymację modeli log-liniowych ze zmiennymi wskaźnikowymi (ang. *response indicator*). Załóżmy, że w analizowanym modelu log-liniowym mamy dwie zmienne wskaźnikowe (*R*, *M*), gdzie *R* wskazuje, czy zmienna *C* jest obserwowalna czy nie, a zmienna wskaźnikowa *M* wskazuje, czy zmienna *D* jest obserwowalna czy nie. Jeśli *C* jest obserwowalna, to *R* przyjmuje wartość 1, w przeciwnym wypadku wartość 2. Jeśli *D* jest obserwowalna, to *M* przyjmuje wartość 1, w przeciwnym wypadku wartość 2.

Vermunt [1997] zmodyfikował algorytm *EM*. Pokazał, jak można zastosować podejście Faya do sytuacji, w której w modelu log-liniowym występują zmienne ukryte. W kroku *E* częstości nieobserwowane tablicy *ABCDRM* są obliczane jako:

$$\begin{aligned}
\hat{n}_{xabcd11} &= n_{abcd}\hat{\pi}_{x/abcd11},\\
\hat{n}_{xabcd12} &= n_{abd}\hat{\pi}_{xc/abd12},\\
\hat{n}_{xabcd21} &= n_{abc}\hat{\pi}_{xd/abc21},\\
\hat{n}_{xabcd22} &= n_{ab}\hat{\pi}_{xcd/ab22}.
\end{aligned} \tag{10}$$

4. Mechanizmy występowania brakujących danych

W pracy [Little, Rubin 1987] zdefiniowano trzy podstawowe mechanizmy występowania brakujących danych: MCAR, MAR oraz NMAR.

MCAR (ang. *missing completely at random)* oznacza, że zmienne wskaźnikowe (*R*, *M*) są zmiennymi niezależnymi w modelu. W tym przypadku nie występuje efekt interakcji pomiędzy tego typu zmiennymi a zmiennymi obserwowalnymi. Tego typu model można zapisać jako:

$$\log(F_{xabcd}) = u + u_x^X + u_a^A + u_b^B + u_c^C + u_d^D + u_r^R + u_s^M + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} + u_{xd}^{XD}. \quad (11)$$

MAR (ang. *missing at random*) oznacza, że w takim modelu log-liniowym, uwzględniającym tego typu mechanizm, występuje efekt interakcji pomiędzy zmiennymi obserwowalnymi a zmiennymi wskaźnikowymi. Kiedy rozważamy podgrupę AB, może wystąpić efekt interakcji, np. zmiennej wskaźnikowej R ze zmienną obserwowalną A oraz B.

$$\log(F_{xabcd}) = u + u_x^X + u_a^A + u_b^B + u_c^C + u_d^D + \sum_j u_j^R + \sum_l u_l^M + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} + u_{xd}^{DX} + \sum_{jl} u_{jl}^{RM} + \sum_{jk} u_{jk}^{YR} + \sum_{jl} u_{jl}^{YM}, \quad (12)$$

gdzie: $Y = A, B, C, D; j, k, l = a, b, c, d$ oraz $j \neq l \neq k$.

NMAR (ang. *non-missing at random*) oznacza, że w modelu log-liniowym będzie rozważany efekt interakcji pomiędzy zmiennymi wskaźnikowymi a zmiennymi, w których występują brakujące dane, np. będzie rozważany efekt interakcji pomiędzy zmienną wskaźnikową R a zmienną D, w której występują brakujące dane.

$$\log(F_{xabcd}) = u + u_x^X + u_a^A + u_b^B + u_c^C + u_d^D + \sum_j u_j^R + \sum_l u_l^M + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} + u_{xd}^{DX} + \sum_{jl} u_{jl}^{RM} + \sum_{jk} u_{jk}^{YR} + \sum_{jl} u_{jl}^{YM}, \quad (13)$$

gdzie: $Y = A, B, C, D; j, k, l = a, b, c, d$ oraz $j = k = l$.

5. Przykład empiryczny

Wykorzystanie analizy klas ukrytych do zmiennych z brakującymi danymi zostało pokazane na przykładzie badania preferencji użytkowników telefonów komórkowych. Badanie zostało przeprowadzone wśród 173 studentów studiów zaocznych w Katowicach. Zostali oni poproszeni o wypełnienie krótkiej ankiety na temat telefonów komórkowych. Znalazły się w niej m.in. pytania dotyczące następujących zagadnień (szczegóły omówiono w tab. 3):

1. Jaki rodzaj telefonu komórkowego Pan/i posiada?
2. Jak często gra w gry wbudowane w komórkę?
3. Czy Pan/i posiada w domu telefon stacjonarny?
4. Czy Pan/i korzysta z MMS-ów?

W tabeli 2 pokazane jest zestawienie miar jakości dopasowania modelu do danych dla modelu uwzględniającego tylko dane pełne oraz modele typu MCAR, MAR oraz NMAR.

Tabela 2. Wyniki statystyk dla dwóch klas ukrytych

Model	χ^2	*df*	L^2	CR	AIC	BIC
Model dla danych pełnych	20,2482	22	17,2478	18,1491	-26,7522	-96,1246
MCAR	55,7260	64	42,4372	47,5485	-85,5628	-290,2666
MAR	31,8569	56	30,3688	29,4098	-81,6312	-260,7470
NMAR	34,5046	54	26,7490	26,7490	-81,2510	-253,9698

Źródło: opracowanie własne z wykorzystaniem programu LEM.

Jak pokazują wartości statystyki χ^2, L^2 oraz CR (statystyka Cressie–Reada) zawarte w tab. 2, wszystkie rozważane modele należy uznać za dobrze dopasowane do danych. Można zauważyć, że ze względu na statystykę BIC model typu NMAR jest modelem lepiej dopasowanym niż model typu MAR.

W przypadku analizy danych pełnych (tab. 3) do klasy pierwszej należy 29,89% respondentów, podczas gdy w drugiej klasie znalazła się większość badanych (70,11%). W klasie pierwszej znalazły się osoby, które korzystają z abonamentu wykupionego w telefonii komórkowej (59,76%), rzadko grają w gry wbudowane w ten telefon (55,84%) oraz 74,93% z nich ma w domu telefon stacjonarny. Natomiast w klasie drugiej znalazły się osoby, które mają prywatny telefon na kartę (62,45%), w ogóle nie grają za pomocą telefonu komórkowego (57,22%) oraz 87,67% z nich ma w domu telefon stacjonarny. W przeciwieństwie do osób, które znalazły się w klasie pierwszej, respondenci należący do tej klasy nie korzystają z MMS-ów (94,48%).

6. Podsumowanie

Na podstawie przeprowadzonego badania ankietowego dotyczącego preferencji użytkowników telefonów komórkowych można stwierdzić, że modele typu MCAR, MAR oraz NMAR dają większy błąd klasyfikacji niż analizowany model dla danych pełnych. Błąd klasyfikacji dla modelu typu MCAR oraz MAR jest taki sam i wynosi 3,97%.

W przypadku modelu dla danych pełnych poprawnie zostało sklasyfikowanych 98,3% respondentów. Zgodnie z miarą redukcji błędu, otrzymane wyniki są o 99,43% lepsze, niż gdybyśmy przydzielili wszystkich respondentów do klasy drugiej.

Tabela 3. Wyniki przeprowadzonej segmentacji dla dwóch klas ukrytych

		Dane pełne		MCAR		MAR		NMAR	
		Klasa I (0,2989)	**Klasa II (0,7011)**	**Klasa I (0,2774)**	**Klasa II (0,7226)**	**Klasa I (0,2759)**	**Klasa II (0,7191)**	**Klasa I (0,2788)**	**Klasa II (0,7062)**
1. Jaki rodzaj telefonu komórkowego Pan/i posiada?	a) prywatny z wykupionym abonamentem	**0,5976**	0,3377	**0,6685**	0,3397	**0,6679**	0,3377	**0,6667**	0,3416
	b) prywatny na kartę	0,3832	**0,6245**	0,3133	**0,6214**	0,3148	**0,6245**	0,3219	**0,6210**
	c) służbowy z wykupionym abonamentem	0,0193	0,0378	0,0183	0,0389	0,0173	0,0378	0,0115	0,0374
2. Jak często gra w gry wbudowane w komórkę?	a) często	0,0579	0,0301	0,0582	0,0312	0,0574	0,0301	0,0380	0,0274
	b) rzadko	**0,5584**	0,3976	**0,6224**	0,3957	**0,6257**	0,3976	**0,6416**	0,3960
	c) w ogóle	0,3836	**0,5722**	0,3194	**0,5731**	0,3169	**0,5722**	0,3204	**0,5766**
3. Czy Pan/i posiada w domu telefon stacjonarny?	a) tak	**0,7493**	**0,8767**	**0,7319**	**0,8767**	**0,7319**	**0,8767**	**0,7386**	**0,8784**
	b) nie	0,2507	0,1233	0,2681	0,1233	0,2681	0,1233	0,2614	0,1216
4. Czy Pan/i korzysta z MMS-ów?	a) tak	**0,9999**	0,0552	**1,0000**	0,0550	**1,0000**	0,0552	**1,0000**	0,0412
	b) nie	0,0001	**0,9448**	0,0000	**0,9450**	0,0000	**0,9448**	0,0000	**0,9588**
	Błąd klasyfikacji	0,017		0,0397		0,0397		0,0291	
	Redukcja błędu (lambda)	0,9943		0,8567		0,8588		0,9009	
	N	173		181		181		181	

Źródło: opracowanie własne z wykorzystaniem programu LEM.

Literatura

Agresti A. (1990), *Categorial Data Analysis*, Wiley, New York.

Fay R.E. (1986), *Causal Models for Patterns of Nonresponse*, „Journal of the American Statistical Association" nr 81, s. 354-365.

Fuchs C. (1982), *Maximum Likelihood Estimation and Model Selection in Coningency Tables with Missing Data*, „Journal of the American Statistical Association" nr 77, s. 270-278.

Goodman L.A. (1974), *Exploratory Latent Structure Analysis Using both Identifiable and Unidentifiable Models*, „Biometrika" nr 61, s. 215-231.

Haberman S.J. (1979), *Analysis of Qualitative Data*, vol. 2, *New Developments*, Academic Press, New York.

Hagenaars J.A. (1990), *Categorical Longitudinal Data: Log-Linear Panel, Trend, and Cohort Analysis*, Sage, Newbury Park.

Kapłon R. (2002), *Analiza danych dyskretnych za pomocą metody LCA*, Taksonomia 9, AE, Wrocław.

Lazarsfeld P.F., Henry N.W. (1968), *Latent Structure Analysis*, Houghton Mill, Boston.

Liao T.F. (2004), *Estimating Household Structure in Ancient China by Using Historical Data: a Latent Class Analysis of Partially Missing Patterns*, „Journal Royal Statistical Society *A*" nr 167, s. 125-139.

Little R.J., Rubin D.B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

McCutcheon A.L. (1987), *Latent Class Analysis*, [w:] *Quantitative Applications in the Social Sciences*, Sage, Newbury Park.

Vermunt J.K. (1997), *Log-Linear Models for Event Histories*, Sage, Thousand Oaks.

LATENT CLASS MODEL FOR CONTINGENCY TABLES WITH MISSING DATA

Summary

The missing data are common problem in many types of data analysis. The main purpose of this paper is to present latent class models in a view of a log-linear approach. This article demonstrates the latent class model for complete data and three basic types of mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR).

This article presents a result of research, which was made among mobile phone users.